Chapter 1

Introduction

Description of:

- Motivation to investigate nearby young open clusters.
- The importance of the IMF
- Current state of knowledge in dynamical simulations.
- The problematic of constrains in dynamical theories.

It must be clear which is the objective

Description of the Nearby open clusters and their properties.

1.1 The DANCe project

- List of open clusters in the DANce project.
- the importance of the pleiades, why we restrict to it.

It must be clear what are the limitations, the boundaries in which the objective will be searched

Description of the current methodologies used to address the question mentioned previously.

- The works of Sarro, Krone-Martins, Malo, Gagne etc.

-The advantages and caveats of the previous methodologies.

-It must be clear the necessity of a new perspective

The proposal we made. The use of Bayesian Hierarchical Models. Benefits and issues of BHM.

Description of the advantages of BHM.

-¿ It must be clear that BHM are the best choice.

Description of the practical issues needed to be solved in order to use BHM.

MCMC techinques and PSO.

-¿ It must be clear that MCMC methods are the best option.

Brief descriptions of our results and how they impact our current knowledge.

-¿ It must be clear that we attained the objective: The pleiades velocity, spatial and mass distributions.

Chapter 2

The Pleiades as a benchmark

2.1 Generalities

Description of the current knowledge on the pleiades cluster Age, distance, metallically reddening, number of members, proper motion. Describe what are they.

2.1.1 The distance controversy

2.2 Spatial Distribution

2.3 Velocity Distribution

2.3.1 Radial Velocity

2.4 Luminosity Distribution

2.5 Mass Distribution

2.6 The current dynamical scenario

2.7 The Pleiades DANCe DR2.

This section must contain a detailed description of the DR2 data.

Table 2.1: Pleiades DANCe DR2 properties

2.7.1 Particularities of the Pleiades DANCe DR2

As described in Section 2.7 the Pleiades DANCe DR2 contains astrometric (stellar positions and proper motions) and photometric ($ugrizYJHK_s$) measurements for 1,972,245 objects.

2.7.2 Selection of variables

Sarro et al. (2014) demonstrated that the most effective variables for the discrimination of members are the proper motions and the $riYJHK_s$ bands. However, they excluded the r band due to its large number of missing values in their training

set. The selection of variables in this work aims at comparing its results with those found by Bouy et al. (2015b) using the $\mu_\alpha, \mu_\delta, J, H, K_s, i-K_s, Y-J$ variables.

The set of variables used in this work are the stellar positions, the proper motions in right ascension and declination, μ_α, μ_δ, and the photometric colours and magnitudes, $i-K_s$,Y, J, H, K_s. However, in order to compare the results with those of Bouy et al. (2015b), the analysis of the spatial distribution of the Pleiades stellar positions is done independently, in Olivares & et al. (2017b).

As described in Olivares & et al. (2017a), the photometry is modelled by parametric series of cubic spline. The parameter of this series is the colour index $i-K_s$ (in the following CI). This colour allows the most one-to-one variate-covariate relation. Figure 2.1 shows the colour-magnitude diagram (CMD) K vs $Y-J$ the second most one-to-one relation.

Figure 2.1: K vs $Y-J$ CMD for the Pleiades candidate members of Bouy et al. (2015b)

.

2.7.3 Data preprocessing

Since both photometry and proper motions carry crucial information for the disentanglement of the cluster population, we restrict the data set to objects with proper motions and at least two observed values in any of our four CMDs: Y, J, H, K_s vs CI. This restriction excludes 22 candidate members of Bouy et al. (2015b), which have only one observed value in the photometry. Furthermore, we restrict the lower limit ($CI = 0.8$) of the colour index to the value of the brightest cluster member. We do not expect to find new bluer members in the bright part of the CMDs. We set the upper limit ($CI = 8$) of the colour index at one magnitude above the colour index of the reddest known cluster member, thus allowing for new discoveries. Due to the sensitivity limits of the DR2 survey in i and K_s bands, objects with $CI > 8$ have K_s magnitudes ≥ 16 mag. These objects are incompatible with the cluster sequence and therefore we discard them a priori as cluster members.

Our current computational constraints and the costly computations associated

to our methodology (described throughout this Sect.), prevent its application to the entire data set. However, since the precision of our methodology, as that of any statistical analysis, increases with the number of independent observations, we find that a size of 10^5 source for our data is a reasonable compromise. Although a smaller data set produces faster results, it also renders a less precise model of the field (in the area around the cluster) and therefore, a more contaminated model of the cluster. For these reasons, we restrict our data set to the 10^5 objects with highest membership probabilities according to Bouy et al. (2015b). Of this resulting data set, the majority ($\approx$98%) are field objects with cluster membership probabilities around zero. Thus, the probability of leaving out a cluster member is negligible. For the remaining of the objects in the Pleiades DANCe DR2, we assign membership probabilities *a posteriori*, once the cluster model is constructed.

Chapter 3

Bayesian formalism

This chapter provides a general introduction to probability theory and its application to parametric inference. Since the objective of this work is to infer the probability distributions of the cluster properties (e.g. luminosity and velocity), I give reason to prove that the Bayes' theorem provides the proper framework for the inference of the parameters governing these distributions. Later in this chapter, I describe the reason for which the Bayesian Hierarchical Models are the best option to parametric inference under the Bayesian framework.

In the following Sections I will describe in detail the assumptions I made to model the data and to select the prior distributions. The two final Sections of this Chapter focus on: the practical issues related to the sampling of the cluster distributions, and the description of details and assumptions embed in the codes I developed.

Partial results of the presented work have been submitted at Olivares & et al. (2017a).

3.1 Introduction to probability theory.

Uncertainty and probability are closely entangled. Anything we measure has an associated uncertainty, otherwise is not a complete measurement [1]. The term uncertainty must not be confused with the term error, which refers to the difference

[1] Upper and lower limits are examples of incomplete measurements.

between the measured value of the quantity and the *true* value[2] of it (for Guides in Metrology 2008). It is commonly agreed that uncertainty of a measurement can be expressed in a probabilistic basis (for Guides in Metrology 2008). This means that whenever we measure a quantity, lets say a, the distribution of the repeated measurements of a, is a probability distribution function, $p(a)$. As any other probability distribution, $p(a)$ satisfies the following properties:

Property 1 It has units, those of the inverse of a.

Property 2 $p(a) \geq 0$.

Property 3 $1 = \int_a p(a)da$.

These properties hold regardless of the dimension of a, it means that the joint uncertainty of all measured quantities of an object is also a probability distribution. Furthermore, they also hold if the probability distribution is conditioned in any other quantity. Lets imagine that we measure the positions, projected in the plane of sky (the plane perpendicular to the line of sight), of one star, these measurements are conditioned in the magnitude (brightness) of the object we measure. If the object is too bright, like the sun, it will saturate the detector and it will render the measurement useless. On the other hand, if the object is too faint we simple will not have enough photons to measure it. So, the stellar positions in the sky, which we can call a and b because they are two dimensions, are conditioned on the magnitude, c, of the object. Therefore, $p(a, b|c)$ must also satisfy:

- It has units of $a^{-1}b^{-1}$.
- $p(a, b|c) \geq 0$.
- $1 = \int_a \int_b p(a, b|c)da \cdot db$.

The link between joint and conditioned probabilities is given by the following symmetric definition:

[2]The true value is that which ideally results when the uncertainty tends to zero.

$$p(a,b) = p(a|b) \cdot p(b).$$
$$p(a,b) = p(b|a) \cdot p(a). \tag{3.1}$$

This can be further conditioned on c to obtain:

$$p(a,b|c) = p(a|b,c) \cdot p(b|c),$$
$$p(a,b|c) = p(b|a,c) \cdot p(a|c), \tag{3.2}$$

If the joint probability of a and b can be factorised, this is

$$p(a,b) = p(a) \cdot p(b),$$
$$p(a,b) = p(b) \cdot p(a), \tag{3.3}$$

then a and b are say to be *independent*. An alternative option is to say that a and b are *independent*, if the conditional probability of a on b is $p(a|b) = p(a)$.

The most important thing we can do with probability distributions is to integrate them. **Property 3** establish that the amount[3] of probability $p(a)$ spread over the volume of the support of a adds to one. This Property allows us to *marginalise* any non-desired variable. Lets imagine again that a and b are the measured positions of some star and we have several measurements of these positions. Then we will have the joint probability distribution of them, $p(a,b)$ (must likely it will be a bivariate gaussian but that does not matter now). If we are interested lets say in the mean value of a, we first must get rid of b. For it, we *marginalise* out b in the following way,

$$p(a) = \int_b p(a,b) \cdot db. \tag{3.4}$$

Then we compute the *expected value* of a, $E(a)$, which is identified with the mean of a once we have drawn many realisations from its probability distribution. To compute it, we add all the possible values of a weighted by their probability. This is,

[3]Which could be infinite, like in Dirac's delta.

$$E(a) = \int_a a \cdot p(a) \cdot da. \tag{3.5}$$

Once again, these last two equations (3.4 and 3.5) hold in case they are conditioned in any other measurement. For example, the magnitude of the object, as in our previous analogy. Notice however that, once the brightness of the object lay within in the dynamic range of the detector, a and b became *independent* of the magnitude.

It is important to recall that the term measurement, and its unavoidable uncertainty, refer not just to directly measured quantities, like the photons (counts) and pixels in a CCD, but also to indirect measurements. Stellar magnitudes and positions in the sky, for example, are indirect measurements derived from the direct measurement of photons, pixels and telescope arrangement. This generalisation also applies to the measurement of parameters in any physical or statistical model, like the one I will describe in the following Section.

3.1.1 Bayes theorem

The definition of conditioned probability (Eq. 3.2) leads to the Bayes' theorem:

$$p(a|b,c) = \frac{p(b|a,c) \cdot p(a|c)}{p(b|c)}. \tag{3.6}$$

Integrating on a we find that,

$$\begin{aligned} p(b|c) \cdot \int_a p(a|b,c) \cdot da &= \int_a p(b|a,c) \cdot p(a|c) \cdot da \\ Z \equiv p(b|c) &= \int_a p(b|a,c) \cdot p(a|c) \cdot da. \end{aligned} \tag{3.7}$$

In this last equation Z refers to what is known as the *evidence*. I will come back to this *evidence* in a few paragraphs. This last Eq. also illustrates that $p(b|c)$ is a normalisation constant which can be evaluated once $p(b|a,c)$ and $p(a|c)$ are known. These two terms are commonly referred as the *likelihood* ($p(b|a,c)$), and the *prior* ($p(a|c)$). Also the term the term $p(a|b,c)$ is called the *posterior*. These names arise

in the context of parametric inference, as I will describe in the next paragraph. However, it worths mention that, although formally the likelihood and the prior are probability distributions in b and a, respectively, for $p(a|b,c)$ to be a probability distribution on a, it only suffices that the product of the likelihood times the prior does not vanish everywhere or be negative anywhere[4]. In this case, they are called *improper* priors or likelihoods. If their product vanishes everywhere, which may be the case if the prior is terribly specified or if the likelihood does not take proper account of extreme data, then the posterior is not a probability distribution due to a division by zero. In any case, it makes no sense try to estimate the parameters of a model with zero evidence.

3.1.1.1 Models and parametric inference

In a broad sense, models are representation or abstraction of the knowledge about something. Sometimes the knowledge is shared by others, some time it is not. They are everywhere in our daily life: from the words we spoke every day, to the evolution of the species and the general relativity; from a kid's draw to the cosmological models. In science, however, we restrict the concept of model to a mathematical representation of the relations among the variables. If the model is parametric, the variables include the data $\mathbf{D}$, which the model attempts to model, and the parameters θ. Parameters are free variables that allow the model to describe the data. Thus a parametric model $\mathcal{M}$, can be represented as:

$$\mathcal{M} = \{f(\mathbf{D}, \theta), \quad \theta \in \Theta\}, \tag{3.8}$$

where f is the function that relates data and parameters, and $\theta \subset \mathbb{R}^k$ with k the dimension of θ. Parametric inference is then the act of finding the distribution of θ given the data $\mathbf{D}$.

The Bayes' theorem allows to perform parametric inference, which is called then Bayesian inference. In this context the Bayes' theorem is:

$$p(\theta|\mathbf{D}, M) = \frac{p(\mathbf{D}|\theta, M) \cdot p(\theta|M)}{p(\mathbf{D}|M)}. \tag{3.9}$$

[4]Although negative probabilities may have sense in quantum mechanics. See for example Dirac (1942)

where $\theta, \mathbf{D}$ and M are correspond, respectively, to the parameters in the model, the data which the model tries to describe, and the prior information used in the construction of the model. Whenever we have a model, we have prior knowledge over it. Actually, it can be classified in two kinds of prior information. One refers to the prior information conveyed in the model, which I call M. This is the information that the creator of the model uses to establish the relations among the elements of the model: variables. The second kind of prior, $p(\theta|M)$ refers to the statement the user of the model made of his/her believes about the probability distribution of the parameter values. This is indeed subjective. However, it is, in my opinion, less subjective than the former, M, prior information. At least in this last kind, the subjectivity is expressed objectively in a probabilistic, and therefore measurable way.

The likelihood of the data $p(\mathbf{D}|\theta, M)$, is a probability distribution on the data, $\mathbf{D}$. However, it is a function on the parameters, θ, which corresponds to the function f of Eq. 3.8. This function is not necessarily a probability distribution on the parameters.

Almost always we assume that the distribution of measurements of one object is independent from the measured value of another object. It means that the collection of measurements we made about for example, the length of a pencil, is independent of the measured length of a pen. This assumption, however, does not always holds. Imagine for example, that we were conducting a statistical analysis on the length of the objects used to define our unit of length. In this case, the statistical distribution of any object is conditioned on the unit of measurement, which in turn is conditioned on the distribution of measurements of the rest of the objects. Therefore, the probability distributions of objects are not independent of themselves[5]. That said, if the data are independent from each other, then

$$p(\mathbf{D}) = \prod_{n=1}^{N} p(d_n), \quad n = \{1, 2, ..., N\} \tag{3.10}$$

with N the number of measurements, and d_n the measurements of a single

[5]This happens for example in satellite surveys for which their system of reference is defined by their own measurements. To avoid this issue their reference systems are anchored on independent measurements.

object. In this case, the joint probability $p(\mathbf{D}|\theta, M)$ can be expressed as,

$$p(\mathbf{D}|\theta, M) = \prod_{n=1}^{N} p(d_n|\theta, M), \quad n = \{1, 2, ..., N\}. \tag{3.11}$$

The term $p(d_n|\theta, M)$ is the likelihood of datum d_n. This term also as called the *generative* model, since it contains the necessary information to generate the data[6].

I interpret the Bayes' theorem, as the probabilistic way to update knowledge. To me, this relation embodies the process of knowledge improvement once we recognise that knowledge is uncertain. In my perspective, knowledge is always uncertain, even if its uncertainty is negligible given the current evidence that supports it. The Bayes' theorem helps us to update our prior knowledge by means of the data (once we multiply them by the likelihood). Then, the posterior probabilities, became our new knowledge. Furthermore, the Bayes' theorem also provides the objective way to compare two models or hypothesis, and update the prior information, M, used to construct them. This is called model selection, which I explain briefly in the next section.

3.1.2 Model Selection

Whenever we have a data set and two or more models that attempt to describe this data, the most straightforward thing to do is to compare these models. Almost always, we want to select the *best* model. Obviously the term *best* depends on the objective of research. For example, lets imagine that our data set consists of a set of bivariate points, for example the measurements of the positions of an object as function of time. If we were interested in reproducing exactly the same points in the data set, the *best* model will be a polynomial with degree equal to the number of points. This polynomial will pass trough all these points. Once we recognise the unavoidable uncertainty of the data, we realise that an exact representation of the data is of poor use. It fits also the noise of the data.

In general, we are interested in the predictive capabilities of a model, its ability to predict future observations rather than to replicate the ones we have. Thus, an

[6]Actually the generative model of the *true* data. To generate the observed data the noise process must be also specified.

exact representation of the observed data (an over-fitted model as in the previous example), will poorly describe any new data set. In this sense, an over-fitted model *memorises* the data rather than *learns* from them.

A model that *learns* from the data is that which obtains the *true* underlying relation embedded in the data. This *true* underlying relation produces the *true* data. The observed data results once the uncertainty is added to it. Nevertheless, we still need to select among different learning models.

We can draw some help from the commonly known Ockham's razor or principle [7]. It says:

Among competing hypotheses, the one with the fewest assumptions should be selected.

Identifying hypothesis with models, this principle tells us we should choose the model with the fewest assumptions. I classify the assumptions of a model in two groups: fixed and free ones. The fixed assumptions belong to what I previously described as the prior information, M, used to construct the model. They may render the model more interpretable in the physically or statistically sense, or even give it coherency within the corpus of a theory. The free assumptions correspond to the parameters of the model. They give it more flexibility to fit the data, although they can also introduce degeneracy in the parametric space. For example, in the case of a straight line model, a fixed assumption is that the data is linearly related, whereas the free assumptions correspond to the slope and ordinate at the origin. Comparing a linear model to a quadratic one in which the constant term has been fixed, we see that they have the same number of free parameters but clearly the second one has an extra fixed assumption. Therefore, choosing the model with fewer free parameters does not necessarily means choosing the model with the fewest assumptions.

One of the great advantages of the Bayesian methodology is that it incorporates directly the Ockham's principle. Suppose that we want to compare two models, M_1 and M_2, which we assume describe the data set $\mathbf{D}$. Each model has prior probabilities, $p(M_k)$ and likelihoods $p(\mathbf{D}|M_k)$ (with $k = 1, 2$). Notice that now, I use the Bayes' theorem for models and not to parameters within a model, as

[7] The origin of this motto and its exact phrasing is beyond the scope of this work. I just mention that paradoxically, an ancient formulation is attributed to Ptolomey: "We consider it a good principle to explain the phenomena by the simplest hypothesis possible" (Franklin 2002)

before. So, the prior probabilities of the models reflect our believes about the fixed assumptions within each model. On the other hand, the likelihood of the data given the model, is related to the parameters (the free assumptions) and priors, both within a model. This likelihood of the data given the model corresponds to the *evidence* of the model, Eq. 3.7. This evidence in terms of the model parameters, θ_k, is now

$$p(\mathbf{D}|M_k) = \int_{\theta_k} p(\mathbf{D}|\theta_k, M_K) \cdot p(\theta_k|M_k) \cdot d\theta_k. \tag{3.12}$$

The Bayes' theorem applied to models instead of individual parameters tells us that

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k) \cdot p(M_k)}{p(\mathbf{D})}. \tag{3.13}$$

with $k = 1, 2$. Since there are only two models, their prior probabilities are related by $p(M_1) = 1 - p(M_2)$. Therefore,

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k) \cdot p(M_k)}{p(\mathbf{D}|M_1) \cdot p(M_1) + p(\mathbf{D}|M_2) \cdot p(M_2)}. \tag{3.14}$$

From this last Equation, the ratio of the posterior distributions is:

$$\frac{p(M_1|\mathbf{D})}{p(M_2|\mathbf{D})} = \frac{p(\mathbf{D}|M_1) \cdot p(M_1)}{p(\mathbf{D}|M_2) \cdot p(M_2)}. \tag{3.15}$$

This ratio provides an objective measure of how better model M_1 is when compared to model M_2, under the measure provided by the data $\mathbf{D}$ by means of the evidence. When both prior probabilities $p(M_1)$ and $p(M_2)$ are set equal, the ratio of posteriors equal the ratio of likelihoods. This is known as the *Bayes factor* (for a similar derivation and some examples of its application see Kass & Raftery 1995). Even in the equal priors case, the evidences themselves, Eq. 3.12, embody the Ockham's principle. Indeed, the evidence is a measure of the prior times the likelihood, this time for parameters in a single model. The larger the number of parameters (assumptions), the larger the volume, in parametric space, over which the likelihood of the data spreads. Since the likelihood is not a probability distribution on the parameters, it does not integrate to one, even if the priors are uniform. The evidence also penalises the assumptions made in the priors of the parameters.

The most concentrated the prior is, the less of the likelihood contributes to the evidence.

Thus, the Bayes' theorem is the way to update knowledge, either if it refers to models or to parameters within a model.

3.1.3 Membership probability

In the previous Section, I derived, by means of the Bayes' theorem, the probability of models M_1 and M_2 given the data $\mathbf{D}$. Now, I describe the same problem but instead of the likelihood of a data set I do it for a single datum. This is, the probability of model M_1 or M_2, given the datum $\mathbf{d}$. This is known as the membership probability of the datum $\mathbf{d}$ to model or class, M_k $(k = 1, 2)$. The Bayes' theorem in this case is,

$$p(M_k|\mathbf{d}) = \frac{p(\mathbf{d}|M_k) \cdot p(M_k)}{\sum_{k=1}^{2} p(\mathbf{d}|M_k) \cdot p(M_k)} \tag{3.16}$$

3.2 Bayesian hierarchical Models

3.2.1 Generalities

Bayesian formalism requires the establishment of priors. As mentioned before, priors represent the *a priori* believe the user of the model has about the possible values that parameters of the model can take. This is indeed subjective. This subjectivity is the main source of criticism from the non-bayesian community [8].

Bayesian hierarchical models, in the following (BHM) can be grouped into the Empirical Bayes methods. In these later ones, the prior distributions are inferred from the data, rather than being directly specified as in common Bayesian methods. In BHM, priors are specified by parametric distributions whose parameters (called hyper-parameters) are also drawn from a parametric distribution in a hierarchical fashion. For this reason, hierarchical models are also called multilevel models. A fully-BHM is that in which the parameters at its higher hierarchy are drawn from a non-parametric distribution. Given its properties, BHM represent the most objective way to the establishment of prior distributions (Gelman 2006).

[8] See Gelman (2012) for a discussion on the ethical use of prior information

Despite the possible high hierarchy of a BHM, for it to be valid, the class of prior distribution must allow the *true* value of its parameter (Morris 1983). For this reason, the updating of knowledge is an important step in the any Bayesian study. If the posterior distribution is in total discrepancy with the prior distribution, or even worst, when the posterior is not fully allowed by the prior distribution (as in a truncated prior for example), then we must update our prior and allow the data to be fully expressed. Otherwise, the posterior could be biased.

Despite its theoretical advantages, BHM are difficult to evaluate since they require far more parameters than standard Bayes methods. Furthermore, their hierarchy (levels) must stop at some point. There are at least two approaches to stop this hierarchy. The first one is to use a non parametric distribution for the hyper-parameters at the higher level. This renders, as previously said, a fully-BHM. However, to use a non-parametric distribution we must have certain prior knowledge about it, which, most of the time is not the case. Another more practical alternative is to give a point estimate, usually the mean or the mode, for the distribution of the hyper-parameter at the top of the hierarchy.

Although in BHM the parameters values of the prior distributions are inferred from data, the user of the model has the important task of specifying the kind of distribution to be used for the prior. Selecting the kind of prior distributions continues to be an active area of research. Common options are include conjugate, non-informative, and weakly informative priors. Conjugate priors are those in which the posterior distribution turns out to be in the same family as the prior distribution, they are called the conjugate of the likelihood. Non-informative and weakly informative priors, as they names indicate, provide intentionally weaker information or no information at all for the prior. Weakly informative priors are the recommended ones (see for example the works of Gelman 2006; Huang & Wand 2013; Chung et al. 2015). Despite the kind of prior distribution chosen, we must always evaluate the prior distribution in terms of the posterior, and check if this last one make sense (Gelman 2006; Gelman et al. 2013, Chap. 6).

3.2.2 Examples

Since BHM usually need more parameters than standard techniques, it restricted its use until modern computers were widely available. Although the idea

of BHM was already present in the 1960s, its application to inference of normal distributions and linear models appears in the 1970s (see Good 1980, for an historical perspective of BHM). In modern days, BHM have a wide range of applications. Some examples of its application are in Gelman & Hill (2007) for the social sciences, Fei-Fei & Perona (2005) for vision recognition and, Diard & Bessiere (2008) for robot navigation.

BHM are widely applied in astrophysics. Although, originally its applications were use mainly in the domain of cosmological parameters inference (see for example the works of Feeney et al. 2013; March et al. 2014; Anderes et al. 2015; Shariff et al. 2016; Alsing et al. 2017), its use was addopted in other domains. Some examples include the study of: the eccentricity distribution of binary stars Hogg et al. (2010b), the Cepheids (Barnes et al. 2004) and RR Lyrae distances (Jefferys et al. 2007), the chemical composition (Wolfgang & Lopez 2015) and albedos of exoplanets (Demory 2014), extinction maps (Sale 2012), stellar parameters (Shkedy et al. 2007), and the present day mass function (Tapiador et al. 2017).

3.2.3 Graphical representation.

Due to the generally large number of parameters in BHM, its interpretation benefits from a graphical representation. Probabilistically Graphical Models (PGM) are graphs that depict the conditional relation among the elements in a model. The elements in a probabilistic model could be constants or stochastic variables. The conditional relations that link elements could be deterministic or stochastic.

In PGM, stochastic variables are represented with circles while constants with squares. If the variable is known, as in the case of the data, it is represented with a filled symbol, otherwise with an empty symbol. Stochastic relations are depicted with solid lines while deterministic ones with dashed lines. If there is no line between two given variables, it indicates that they are assumed to be independent. Variables that repeat together, as in the case of the data, are grouped within a plate. The number of repetitions is indicated in one corner of the plate. The community generally agrees in these set of standard representations (for more details on PGM see for example the book of Koller & Friedman 2009). Figure 3.1 shows a simple example of a PGM for the inference of the parameters of a normal

distribution.

Figure 3.1: PGM representing the parametric inference of a normal distribution.

3.3 Modelling the data

Creating a model is a complex task. As previously mentioned, a model is the mathematical representation of the knowledge about something. In this work my objective is the modelling of the DANCe data related to NYOC (nearby young open clusters). Since my aim is the statistical description of the NYOC population, most of the time, the relations I use are statistical.

Modelling a data set demands the gathering and sorting of the prior knowledge. This last refers to the knowledge about the data set, the object of study, the statistical techniques that may help to attain the objective, and the computational resources at hand. I collected this knowledge from three main sources: the standard references (e.g. articles and books), my colleagues and experts (*knowledge elicitation*), and my self. Arraigning the prior knowledge into the model is an iterative and therefore continuos process. In thisSection I describe an snapshot of this process. The state of the model once the article Olivares & et al. (2017a) was submitted. Later in Section 3.6 I will give a brief description of the model development process in the context of its coded versions.

In this Section I will describe one crucial aspect of the DANCe DR2 data set: the missing values. Then I will describe the relevant knowledge of NYOC and how I embedded this knowledge in the the data model in terms of field and cluster models. Finally, I will describe the details of the prior distributions and of its hierarchy.

In the following, whenever I use the pronoun *we*, it refers to the authors of Olivares & et al. (2017a), where a synthesis of this work is presented. I use the pronoun I to emphasis that the particular idea or task was done by me.

3.3.1 Missing values

Missing values can happen due to different processes. From the physical perspective, they can arise due to faint or bright sources that produce counts which are outside the dynamical range of the detector. They can also emerge due to for example detector or random issues (e.g. electronic failures or cosmic rays). From the statistical perspective however, the most important aspects of missing values are the probability distribution of their occurrence, and the fact that they are partially or completely missing. A partially missing value (or partially observed) is that for which an upper or lower limit is given whereas a completely missing value is simply not available at all. The DANCe survey contains only (so far) completely missing values. Upper and lower limits could also be inferred from the data provided that missing values occur outside these limits. However, I leave aside this task since missing values in DANCe data occur also at the interior of the variables domain.

In terms of probability, there is no distinction between missing values and parameters. Therefore, we can marginalise missing values as we do with any other nuisance parameter. If datum $\mathbf{d}$ has a missing entry, $\mathbf{d} = \{d_1, d_2, ..., mis, ..., d_n\}$, with n the dimension of $\mathbf{d}$, then, the likelihood of this datum, given model parameters θ is

$$p(\mathbf{d}|\theta) = \int_{-\infty}^{\infty} p(\{d_1, d_2, ..., mis, ..., d_n\}|\theta) dmis. \tag{3.17}$$

Throughout this work, missing values are marginalised in this way.

I made a remark of a point that may seem obvious but it is important to remember. Let $p(a)$ be a probability distribution, A a random sample of n point from it, and $p_A(a)$ the empirical probability distribution of A. Let B be a non-random sample of $p(a)$ with n elements, and empirical probability distribution $p_B(a)$. In the limit of $n \to \infty$, $p_A(a) = p(a)$ however $p_B(a) \neq p(a)$. Therefore, $p_A(a) \neq p_B(a)$. Similarly, in data with missing values, if the missing value pattern is not random, the distribution of the completely observed data (with non-missing values) differs from that of all the data. In Olivares & et al. (2017a) we show this subtle but important difference in the case of the DANCe data set.

3.3.2 The generative model

Since the objective of this work is the statistical study of NYOC we must separate them from the field population. To perform this separation, we use the data, which as always, is uncertain. Therefore, the separation is also uncertain. As mentioned in Chapter **??**, under the current set of variables, the cluster and the field are entangled. To probabilistically disentangle them, we must provide probabilistic models for both populations. These models are the likelihoods. With these likelihoods and prior probabilities for the cluster and field model, we are able to compute the membership probability (Eq. 3.16) of each object in our data set to the cluster model. This model is assumed to represent the cluster population. First, this model must be learnt from the data.

The learning process demands a set of N binary integers $\mathbf{q}$, one q_n for each object. Each of these two possible values represent one of the two mutually exclusive possibilities: the object belongs to the cluster ($q_n = 1$) or to the field population ($q_n = 0$). Let θ and p_c be the parameters and model of the cluster. Also, ϕ and p_f the parameters and model of the field. Then, the likelihood of the data is,

$$p(\mathbf{D}|\mathbf{q}, \theta, \phi) = \prod_{n=1}^{N} p_c(\mathbf{d}_n|\theta)^{q_n} \cdot p_f(\mathbf{d}_n|\phi)^{(1-q_n)}. \tag{3.18}$$

This $\mathbf{q}$ is now marginalised using a probability for it, a prior probability or measure which is set in terms of a new and unique parameter π, which represent the *prior* probability that an object belongs to the field. Thus, the probability of $\mathbf{q}$ is

$$p(\mathbf{q}|\pi) = \prod_{n=1}^{N} (1-\pi)^{q_n} \cdot \pi^{(1-q_n)}. \tag{3.19}$$

and the marginalisation runs as

$$
\begin{aligned}
p(\mathbf{D}|\pi,\theta,\phi) &= \int_{\mathbf{q}} p(\mathbf{D},\mathbf{q}|\pi,\theta,\phi)\cdot d\mathbf{q} \\
&= \int_{\mathbf{q}} p(\mathbf{D}|\mathbf{q},\pi,\theta,\phi)\cdot p(\mathbf{q}|\pi)\cdot d\mathbf{q} \\
&= \int_{\mathbf{q}} \prod_{n=1}^{N} p_c(\mathbf{d}_n|\theta)^{q_n}\cdot p_f(\mathbf{d}_n|\phi)^{(1-q_n)}\cdot \prod_{n=1}^{N}(1-\pi)^{q_n}\cdot \pi^{(1-q_n)}\cdot d\mathbf{q} \\
&= \int_{\mathbf{q}} \prod_{n=1}^{N} \left[(1-\pi)\cdot p_c(\mathbf{d}_n|\theta)\right]^{q_n}\cdot \left[\pi\cdot p_f(\mathbf{d}_n|\phi)\right]^{(1-q_n)}\cdot d\mathbf{q} \\
&= \prod_{n=1}^{N}(1-\pi)\cdot p_c(\mathbf{d}_n|\theta)+\pi\cdot p_f(\mathbf{d}_n|\phi). \qquad (3.20)
\end{aligned}
$$

This last equality is a rather complicated derivation which can be found on Press (1997); Hogg et al. (2010a) for p_c and p_f in the exponential family. Also, a general derivation of this expression is given by Jaynes (2003). He obtains it assuming individual unknown probabilities p_n instead of q_n and marginalising over them by the aid of a prior.

Thus, the *generative model* or likelihood of the datum $\mathbf{d_n}$ is

$$
p(\mathbf{d}_n|\pi,\boldsymbol{\theta}_c,\boldsymbol{\theta}_f,\mathbf{u}_n) = \pi\cdot p_f(\mathbf{d}_n|\boldsymbol{\theta}_f,\mathbf{u}_n)+(1-\pi)\cdot p_c(\mathbf{d}_n|\boldsymbol{\theta}_c,\mathbf{u}_n), \qquad (3.21)
$$

where $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_c$ indicate the cluster and field parameters, while $\mathbf{u}_n$ refers to the datum uncertainty. The probabilities $p_f(\mathbf{d}_n|\boldsymbol{\theta}_f,\mathbf{u}_n)$ and $p_c(\mathbf{d}_n|\boldsymbol{\theta}_c,\mathbf{u}_n)$ are the field and cluster models, respectively. These models are explained in detail in the next two sections.

In the following, I assume that the observed quantities, even if they contain missing values, resulted from the convolution[9] of the *true* quantities with a source of uncertainty. If I were to assume that uncertainties of individual objects were drawn from the same uncertainty distribution, then these objects will be independent and identically distributed (commonly known as i.i.d). However, this

[9]The addition of two stochastic variables is analogous to the convolution of their probability distributions.

assumption is not necessary. Instead, we model the data with its intrinsic heteroscedasticity. It means that individual observations have different dispersions. However, we assume that the multivariate normal is the family distribution for these uncertainties. This assumption is standard practice and is also supported by the large and heterogeneous origins of the DANCe data.

3.3.3 The field population

To model the field population, we assume that the joint probability distribution of the data can be factorised into the probability distributions of proper motion and photometry. Thus, they are independent, at least conditioned on the parameters. It is also assumed that both distributions are described by Gaussian Mixture Models (GMM). The flexibility of GMM to fit a variety of probability distribution geometries make them a suitable model to describe the density of the heterogeneous data from the DANCe DR2.

A GMM is a probability distribution resulting from the linear combination of M gaussian distributions, $\pi_m \cdot \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, with $m = 1, 2, ..., M$. Where π_m are the amplitudes or fractions, which must add to one. Thus

$$p_{GMM}(x|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{m=1}^{M} \pi_m \cdot \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{3.22}$$

Notice that the number of gaussians in the mixture is no formally speaking a parameter, but rather it implies a collection of parameters. The number of these parameters increases quadratically with the number of gaussians.

According to Bouy et al. (2015b), the number of Pleiades candidate members in their data set is 2109. This means that the number of field objects dominates. Thus, it can be assumed that any reclassification of candidate members will have a negligible impact on this figures. Therefore, it seems reasonable to assume that the GMM describing the field population can be fixed during the process of cluster parameters inference. Let me elaborate on this assumption.

The field objects of Bouy et al. (2015b) are those whose cluster membership probability is lower than 0.75. There are $\sim 98,000$ of these objects. These authors selected this probability threshold based on the analysis that Sarro et al. (2014) made of their methodology when applied to synthetic data. Sarro et al.

(2014) report, at probability threshold $p = 0.75$ contamination and true positive rates of $\approx 8\%$ and $\approx 96\%$ respectively. Therefore the number of hypothetically misclassified objects,12% of the cluster members, ≈ 258, is negligible compared to the size of the data set ($100,000$ objects, the restricted sample). It represents a negligible fraction ($\lesssim 0.26\%$). Furthermore, under the assumption that the work of Sarro et al. (2014) is correct, the miss classified objects concentrate on cluster membership probabilities near the classification threshold. Therefore, they will also group in an area of the physical "boundary" between the cluster and the field (I call it physical because it is on the observable variables and not in the probability threshold)). This is the area of highest entanglement. It does not meant that misclassified objects will not lay in the core of the cluster (objects with high membership probabilities). It means that the occurrence of these cases will be lower. This physical "boundary" will correspond, in proper motions space to a halo around the cluster centre. In photometry, however, this boundary will run all along the cluster sequence in the CMDs. All previous assumption are there to justify that the negligible fraction of hypothetical misclassified is not concentrated in the physical space.

Therefore, if the misclassified objects are a few and spread over the space, their contribution to the parameters of the GMM describing the field population can be neglected. Thus, the parameters of the GMM remain fixed and out of the inference process.

The number of gaussians in each GMM was found using the Bayesian Information criterion (BIC, Schwarz 1978). The BIC is a model selection criteria that aims at avoiding over fitting. It represents a compromise between the likelihood, $\mathcal{L}$, of the n data points, and the number of parameters, k. This is,

$$BIC = \ln n \cdot k - 2 \ln \mathcal{L} \tag{3.23}$$

To estimate the parameters of the GMM that maximise the likelihood, we used the Expectation Maximisation (EM) algorithm. However, the missing values in the photometry prevent the use of the standard from of the algorithm (see for example Chapter 9 of Bishop 2006). Instead, the parameters were estimated with the modified version of the EM for GMM of McMichael (1996). On it, objects with missing values also contribute to estimate the maximum-likelihood (ML) parameters. The

number of gaussians suggested by the BIC for this mixture is 14.

Regarding the proper motions, the standard EM fro GMM can be used. However, the BIC favours models in which the number of gaussians is large, their fractions small and their variances also large. To circumvent this issue, a uniform distribution was added to the GMM. The EM algorithm was modified accordingly. The BIC applied to this new mixture of distributions renders reasonable results. This modification improves the likelihood while reduces the number of parameters. The number of gaussians suggested by the BIC for this mixture is 7, plus the uniform distribution.

Thus, the field likelihood $p_f(\mathbf{d}|\boldsymbol{\theta}_f, \mathbf{u})$ of an object with measurements $\mathbf{d}$, given parameters, $\boldsymbol{\theta}_f$, and standard uncertainties $\mathbf{u}$ is

$$\begin{aligned} p_f(\mathbf{d}|\boldsymbol{\theta}_f, \mathbf{u}) = [&c \cdot \pi_{f,pm,0} + \\ &\sum_{i=1}^{7} \pi_{f,pm,i} \cdot \mathcal{N}(\mathbf{d}_{pm}|\boldsymbol{\mu}_{f,pm,i}, \boldsymbol{\Sigma}_{f,pm,i} + \mathbf{u}_{pm})\Bigg] \\ &\cdot \left[\sum_{i=1}^{14} \pi_{f,ph,i} \cdot \mathcal{N}(\mathbf{d}_{ph}|\boldsymbol{\mu}_{f,ph,i}, \boldsymbol{\Sigma}_{f,ph,i} + \mathbf{u}_{ph})\right]. \end{aligned} \tag{3.24}$$

In this equation, $\boldsymbol{\theta}_f$ refers to the set of field parameters, $\boldsymbol{\pi}_f, \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f$, which are, respectively, the fractions, means and covariance matrices of the GMM. The first and second brackets represent the proper motion and photometric models, respectively. The first term of the proper motion model is the uniform distribution. In it, c is a constant determined by the inverse of the product of the proper motions ranges (see Table 2.1), and $\pi_{f,pm,0}$ is the fraction of this uniform distribution. The second term in the same bracket is the mixture of gaussians with means $\boldsymbol{\mu}_{f,pm}$ and covariance matrices $\boldsymbol{\Sigma}_{f,pm} + \mathbf{u}_{pm}$.

I refer the interested reader to Appendix **??**, which contains specific details of the GMM presented in this section.

3.3.4 The cluster population

Similarly to what it was assumed for the field population, for the cluster population we assumed also that its data distribution can be factorised in the product of the proper motions distributions times the photometric distribution. It is known

that unresolved systems of stars (groups of stars that, given the spatial resolution of the telescope, are seen as an individual object) have an increased brightness proportional to the multiplicity of the system. In particular, if an unresolved system is made of two equally luminous objects, then its magnitude is 0.752 times brighter than that of an individual object. This is the case of an equal mass binary.

Sarro et al. (2014) show evidence of an equal-mass binaries (EMB) sequence in the Pleiades cluster. Those authors model this parallel sequence assuming that the number of objects in this sequence is 20% of the total number of members. In this work, we also model this EMB sequence. However, we do not assume its proportion and rather infer it from the data.

Unresolved multiple systems have an impact on proper motion. From an astrophysical point of view, in stellar clusters, massive objects fall towards the centre of the gravitational potential in a higher rate than less massive ones due to stellar encounters (reference needed). From an astronomical point of view, an unresolved binary system shifts the photo centre of its images when compared to that of a single object. For the previous reasons, we model the EMB as an independent population in the proper motions, and pair this model to its photometric parallel 0.75 displaced model. This statistical model allows a comparison of the EBM population with that of the rest of the cluster. We call all non EMB objects single stars. Although this is an abuse of the terminology because there are binaries and multiple systems with different mass ratios, it keeps the text more readable.

3.3.4.1 Photometric model of equal-mass binaries and single stars

To model the cluster sequence in the CMDs, both for single and EMB, we use cubic splines for each of the $YJHK_s$ vs CI CMDs. This assumption roots on the fitting properties of splines. I tried several polynomial bases (Laguerre, Hermite, Chebyshev) but no matter the order, they were not able to fit the sequence, particularly in the region around $CI \approx 3$, where the slope is higher. Despite their flexibility, these splines require more parameters than common polynomials. If represented in terms of B-splines series, in addition to the coefficients of the series, they require the specification of points called knots. These knots represent the starting and ending points of the spline sections. By definition, a basis spline (B-spline) of order n is a piece wise polynomial function of order n in the interval

$t_0 \leq x \leq t_n$. The boundary and internal points $\mathbf{t}$ are called knots. In particular, for a given set of knots, there is only one B-spline, thus the name basis spline. Thus, any spline function of order n can be represented as a series of b-splines. In particular, any cubic spline can be represented as,

$$S_3(CI, \boldsymbol{\beta}, \mathbf{t}) = \sum_i \beta_i \cdot B_{i,3}(CI, \mathbf{t}). \tag{3.25}$$

Where $B_{i,n}$ are given by the Cox-de Boor recursive formula and $\boldsymbol{\beta}$ the coefficients of the series. For more details on splines and the Cox-de Boor formula see De Boor (1978).

Despite their fitting properties, B-splines have an issue when inferring simultaneously their coefficients and knots: there is multi modality in the parametric space (Lindstrom 1999). It means that at least more than one combination of parameters produces the same solution. To avoid this multi modality, I decided to keep the knots fixed throughout inference. This decision, although reduces the flexibility of the splines, allows a still better fit than common polynomials. To obtain the ML estimate of the knots I use the algorithm of Spiriti et al. (2013). This algorithm, implemented in the *freeknotsplines* R package, allows to simultaneously obtain the best truncation value for the spline series. It uses the BIC to select among different models. In order to obtain these figures, I used the candidate members of Bouy et al. (2015b). The BIC indicates that seven coefficients is the best number with knots at $\mathbf{t} = \{0.8, 3.22, 3.22, 5.17, 8.0\}$. I tested different number of knots, ranging from two to nine, with five the best configuration given by the BIC.

As I mentioned in the introduction to this Section, we assume that the observed photometric quantities are drawn from a probability distribution resulting from the convolution of the observed uncertainties, with the *true* quantities. Here comes an expiation. I recognise that the model is far from perfect and that there are several astrophysical phenomena that does not address. Either because I do not have the knowledge or because they are too complicated to model. These phenomena include but are not limited to age, metallicity and distance dispersions, unresolved systems (other than EMB), variability, transits, etc. So, instead of modelling them I assume that all of them, even the unknown, contribute to an intrinsic photomet-

ric dispersion. Given the large and unknown sources contributing to this intrinsic dispersion, we can safely assume that it is multivariate gaussian, whose parameters we infer from the data. We assume that it describes the intrinsic dispersion of both cluster and EMB sequences. But most importantly, we distinguish this dispersion from the photometric uncertainty of individual measurements. If we were to assume no *true* intrinsic dispersion, then any deviation from the *true* quantities should have to be explained *only* by the observational uncertainty.

In this multivariate gaussian modelling the intrinsic dispersion. The splines model the mean of the *true* photometric quantities, both for the cluster sequence, $\boldsymbol{t}_{ph;Cs}$, and the EMB, $\boldsymbol{t}_{ph;Bs}$. The covariance matrix, Σ_{clus}, represents the width of the intrinsic dispersion. Covariance matrices are symmetric and positive semi-definite. Therefore, from the 25 entries in this 5×5 matrix, only 15 are different. We use the Cholesky decomposition to obtain a triangular matrix whose parameters we infer from the data.

Thus, the mean *true* photometry is

$$\begin{aligned}\boldsymbol{t}_{ph;Cs} &= \{CI, Y, J, H, K_s\} \\ \boldsymbol{t}_{ph;Bs} &= \{CI, Y - 0.75, J - 0.75, H - 0.75, K_s - 0.75\}\end{aligned}$$

where

$$\begin{aligned}Y &= \mathcal{S}_Y(CI, \beta_Y) \\ J &= \mathcal{S}_J(CI, \beta_J) \\ H &= \mathcal{S}_H(CI, \beta_H) \\ K_s &= \mathcal{S}_{K_s}(CI, \beta_{K_s})\end{aligned}$$

and β_{Y,J,H,K_s} denote the coefficients of all the splines (a 4×7 matrix),by simplicity I call it $\boldsymbol{\beta}$.

Since the photometry of the EMB is a linear transformation, T_{Bs}, of the mean *true* photometry of cluster sequence, no extra parameters are required. Therefore,

$$\boldsymbol{t}_{ph;Cs} = \boldsymbol{\mathcal{S}}(CI, \boldsymbol{\beta}) \tag{3.26}$$

$$\boldsymbol{t}_{ph;Bs} = T_{Bs}(\boldsymbol{\mathcal{S}}(CI, \boldsymbol{\beta})). \tag{3.27}$$

The cluster and EMB likelihoods of an object with photometric measurements $\mathbf{d}_{ph}$ and standard uncertainties $\mathbf{u}_{ph}$ are:

$$\begin{aligned} p_{Cs}(\mathbf{d}_{ph}|CI, \boldsymbol{\beta}, \Sigma_{clus}, \mathbf{u}_{ph}) &= \mathcal{N}(\mathbf{d}_{ph}|\boldsymbol{t}_{ph;Cs}, \mathbf{u}_{ph} + \Sigma_{clus}), \\ p_{Bs}(\mathbf{d}_{ph}|CI, \boldsymbol{\beta}, \Sigma_{clus}, \mathbf{u}_{ph}) &= \mathcal{N}(\mathbf{d}_{ph}|\boldsymbol{t}_{ph;Bs}, \mathbf{u}_{ph} + \Sigma_{clus}), \end{aligned} \tag{3.28}$$

where $\boldsymbol{t}_{ph;Cs}$ and $\boldsymbol{t}_{ph;Bs}$ are given by Equations 3.26 and 3.27, respectively.

Since the splines are parametrised by the true CI of each object, we have more parameters than objects [10]. This *true* CI is unknown even if its observed value is not missing. We solve this problem (it is a computational problem!) by marginalising these nuisance parameters (See Eq. 3.30 and 3.31). To marginalise these CI we need a measure or prior. We provide a prior, in a hierarchical way, (thus the name Bayesian Hierarchical model) and then marginalise the parameter. This marginalisation lefts behind a precise estimate of the parameters of the prior distribution. To this estimate all objects contributed, paradoxically also those with a missing CI. This is the force of the BHM.

We model the prior for the *true* CI as a truncated ($0.8 \leq CI \leq 8$) univariate GMM with five components whose parameters are also inferred from the data. We chose five components as BIC suggested. I applied BIC to the results found using the EM algorithm and the candidate members of Bouy et al. (2015b). I tested higher values but the posterior distribution did not changed much, thus I preferred the BIC value.

This GMM is

$$p_{CI}(CI|\boldsymbol{\pi}_{CI}, \boldsymbol{\mu}_{CI}, \boldsymbol{\sigma}_{CI}) = \sum_{i=1}^{5} \pi_{CI,i} \cdot \mathcal{N}_t(CI|\mu_{CI,i}, \sigma_{CI,i}). \tag{3.29}$$

In this last Equation, the symbol $\mathcal{N}_t$ stands for the truncated ($0.8 < CI < 8$)

[10] Although this sounds crazy, the rules of probability calculus do not discard this possibility.

univariate normal distribution.

Then, the marginalisation of CI runs as follows:

$$\begin{aligned} p_{Cs}(\mathbf{d}_{ph}|\boldsymbol{\theta}_c, \mathbf{u}_{ph}) &= \int p_{Cs}(\mathbf{d}_{ph}, CI|\boldsymbol{\theta}_c, \mathbf{u}_{ph}) \cdot dCI \\ &= \int p_{Cs}(\mathbf{d}_{ph}|CI, \boldsymbol{\theta}_c, \mathbf{u}_{ph}) \cdot p_{Cs}(CI|\boldsymbol{\theta}_c, \mathbf{u}_{ph}) \cdot dCI \end{aligned} \tag{3.30}$$

$$\begin{aligned} p_{Bs}(\mathbf{d}_{ph}|\boldsymbol{\theta}_c, \mathbf{u}_{ph}) &= \int p_{Bs}(\mathbf{d}_{ph}, CI|\boldsymbol{\theta}_c, \mathbf{u}_{ph}) \cdot dCI \\ &= \int p_{Bs}(\mathbf{d}_{ph}|CI, \boldsymbol{\theta}_c, \mathbf{u}_{ph}) \cdot p_{Bs}(CI|\boldsymbol{\theta}_c, \mathbf{u}_{ph}) \cdot dCI. \end{aligned} \tag{3.31}$$

In these Equations, θ_c stands for all cluster parameters related to photometry, and the first and second terms of the integrals in the last equalities correspond to Equations 3.28 and 3.29, respectively. The distribution of CI depends only on $\boldsymbol{\pi}_{CI}, \boldsymbol{\mu}_{CI}, \boldsymbol{\sigma}_{CI}$, thus, the cluster and equal-mass binaries likelihoods of datum $\mathbf{d}_{ph}$ are

$$\begin{aligned} &p_{Cs}(\mathbf{d}_{ph}|\boldsymbol{\pi}_{CI}, \boldsymbol{\mu}_{CI}, \boldsymbol{\sigma}_{CI}, \boldsymbol{\beta}, \Sigma_{clus}, \mathbf{u}_{ph}) \\ &= \int \mathcal{N}(\mathbf{d}_{ph}|\boldsymbol{\mathcal{S}}(CI, \boldsymbol{\beta}), \mathbf{u}_{ph} + \Sigma_{clus}) \\ &\cdot \sum_{i=1}^{5} \pi_{CI,i} \cdot \mathcal{N}_t(CI|\mu_{CI,i}, \sigma_{CI,i}) \cdot dCI \\ &p_{Bs}(\mathbf{d}_{ph}|\boldsymbol{\pi}_{CI}, \boldsymbol{\mu}_{CI}, \boldsymbol{\sigma}_{CI}, \boldsymbol{\beta}, \Sigma_{clus}, \mathbf{u}_{ph}) \\ &= \int \mathcal{N}(\mathbf{d}_{ph}|T_{Bs}(\boldsymbol{\mathcal{S}}(CI, \boldsymbol{\beta})), \mathbf{u}_{ph} + \Sigma_{clus}) \\ &\cdot \sum_{i=1}^{5} \pi_{CI,i} \cdot \mathcal{N}_t(CI|\mu_{CI,i}, \sigma_{CI,i}) \cdot dCI. \end{aligned} \tag{3.32}$$

The observed CI and magnitudes help us to reduce the computing time of the marginalisation integral. We use them to discard regions of the marginalisation integral in which the argument is almost zero (i.e. far from the measured values). Although we allow the nuisance parameter to have all its possible values, the data, through the likelihood, gives us information about the parameter distribution.

To use this information, we proceed as follows: first, we compare the observed photometry to the true one (i.e. the cluster sequence given by the splines) in a grid of 300 points uniformly distributed in the domain of CI, then, we find the point, p, from the grid which is the closest to the observed photometry. To compute distance we use the Mahalanobis metric. This metric takes into account the observational uncertainty $\mathbf{u}$ and the intrinsic dispersion of the cluster sequence, Σ_{clus}. To define the limits of the marginalisation integral, we use a ball of 3.5 Mahalanobis distances around point p. Contributions outside this ball are negligible ($< 4 \times 10^{-4}$).

3.3.4.2 Proper motion model of equal-mass binaries and single stars

As mentioned before, we assume that the cluster population has two, mutually exclusive, groups: single and EMB stars. We model the proper motions of these two subpopulations with GMM, one model for each population. If the cluster is virialised (see Chapter 2), we can assume that is velocity distribution is almost Maxwellian (Maxwell-Boltzman distribution). Therefore a GMM is a reasonable assumption. Furthermore, in a virialised system we expect radial symmetry. Thus we assume that the gaussians within the GMM share the same mean, However we allow a independent means for single and EMB GMM. The assumption of radial symmetry may be a weak one in the presence of the galactic potential. The galactic potential can perturb the cluster and deviate its velocity distribution from radial symmetry. Nevertheless, since we model the covariance matrix of the GMM as full covariance matrices, departures from the radial symmetry can still be perfectly model by the different eigen-vectors of these matrices.

We infer the parameter of these GMM as part of our Bayesian hierarchical model. However we set a priori the number of gaussians in each GMM. Not doing so will demand a technique in which the model parameters can be augmented. Although such techniques already exit, they are still under computational development (see Fan & Sisson 2011, for a review of reversible jump MCMC).

Using the EM for GMM and the proper motions of the candidate members of Bouy et al. (2015b) I obtained the ML estimates for the GMM likelihood. I did this for configurations of GMM ranging from one to five components. Using the BIC criterium (Eq. 3.23), I selected four and two gaussians for single and EMB GMM, respectively. Since covariance matrices are always symmetric, only three

parameters are needed to fully specify the covariance matrix of these bivariate normal distributions.

Thus, the cluster (subindex Cs) and EMB (subindex Bs) likelihoods of an object with proper motions measurements $\mathbf{d}_{pm}$ and uncertainties $\mathbf{u}_{pm}$ are

$$p_{Cs}(\mathbf{d}_{pm}|\boldsymbol{\pi}_{Cs}, \boldsymbol{\mu}_{Cs}, \boldsymbol{\Sigma}_{Cs}, \mathbf{u}_{pm}) = \sum_{i=1}^{4} \pi_{Cs,i} \cdot \mathcal{N}(\mathbf{d}_{pm}|\boldsymbol{\mu}_{Cs}, \Sigma_{Cs,i} + \mathbf{u}_{pm})$$
$$p_{Bs}(\mathbf{d}_{pm}|\boldsymbol{\pi}_{Bs}, \boldsymbol{\mu}_{Bs}, \boldsymbol{\Sigma}_{Bs}, \mathbf{u}_{pm}) = \sum_{i=1}^{2} \pi_{Bs,i} \cdot \mathcal{N}(\mathbf{d}_{pm}|\boldsymbol{\mu}_{Bs}, \Sigma_{Bs,i} + \mathbf{u}_{pm}). \quad (3.33)$$

Finally, combining the proper motions and photometric models, the total cluster likelihood of an object with measurement $\mathbf{d}$ and uncertainties $\mathbf{u}$ is

$$\begin{aligned} p_c(\mathbf{d}|\boldsymbol{\theta}_c, \mathbf{u}) = \pi_{CB} &\cdot p_{Cs}(\mathbf{d}_{pm}|\boldsymbol{\pi}_{Cs}, \boldsymbol{\mu}_{Cs}, \boldsymbol{\Sigma}_{Cs}, \mathbf{u}_{pm}) \\ &\cdot p_{Cs}(\mathbf{d}_{ph}|\boldsymbol{\pi}_{CI}, \boldsymbol{\mu}_{CI}, \boldsymbol{\sigma}_{CI}, \boldsymbol{\beta}, \Sigma_{clus}, \mathbf{u}_{ph}) \\ &+ (1 - \pi_{CB}) \cdot p_{Bs}(\mathbf{d}_{pm}|\boldsymbol{\pi}_{Bs}, \boldsymbol{\mu}_{Bs}, \boldsymbol{\Sigma}_{Bs}, \mathbf{u}_{pm}) \\ &\cdot p_{Bs}(\mathbf{d}_{ph}|\boldsymbol{\pi}_{CI}, \boldsymbol{\mu}_{CI}, \boldsymbol{\sigma}_{CI}, \boldsymbol{\beta}, \Sigma_{clus}, \mathbf{u}_{ph}), \end{aligned} \quad (3.34)$$

where π_{CB} is the parameter representing the amplitude of single cluster sequence stars in the single-EMB mixture model. The photometric and proper motions likelihoods are given by Equations 3.32 and 3.33, respectively.

3.4 Priors

The Bayesian formalism is characterised by the use of priors. These priors, as mentioned earlier, represent the objective and measurable way to establish believes about the distribution of values that certain parameter may have. Although the formal way to establish believes in the form of a probability distribution is objective[11]. The believes themselves remain subjective (influenced by the sentience of the subject).

[11]It is reproducible and measurable

In the following I describe the information used to establish both the family of the prior distribution as well as its hyperparameters. The priors we assume are intended to fall, in most of the cases, in the category of weakly informative priors. A weakly informative prior provides weaker information than the one actually available Gelman (2006). This kind of priors show better computational performance when compared to non-informative priors. Examples of these can be found in the works of Gelman et al. (2008) and Chung et al. (2015).

I group priors into three main categories. Fractions in general, and those concerning: proper motions and photometric models.

Fractions are defined for mixtures. The mixtures in the model are the GMMs, the cluster-field, and the single-EMB mixtures. At each mixture, the fractions or amplitudes quantify the contribution of each element to the mixture. If each element in the mixture is itself a probability distribution, then the fractions must add to one and be bounded by the $[0, 1]$ interval. We choose the Dirichlet distribution since it is the multivariate generalisation of the beta distribution. It has support in $[0, 1]$ for each entry, and is parametrised by $\boldsymbol{\alpha}$. This distribution gives the probability of n rival events given that each rival event has been observed $\alpha_i - 1$ times $(i = \{1, 2, ..., n\})$.

The mean and variance of the Dirichlet distribution are given by,

$$E[x_k] = \frac{\alpha_k}{\sum_k \alpha_k}, \tag{3.35}$$

$$Var[x_k] = \frac{-\alpha_k \cdot (\alpha_k - \sum_k \alpha_k)}{(\sum_k \alpha_k)^2 \cdot (1 + \sum_k \alpha_k)}. \tag{3.36}$$

For the field-cluster mixture we set the hyper-parameters to $\boldsymbol{\alpha} = \{9.8, 0.2\}$. We expect a mean 98% of field objects and a 2% of cluster objects with little variance. These figures correspond to the fraction of field and cluster candidate members found by Bouy et al. (2015b). For the single-EMB mixture we use an hyper-parameter value, $\boldsymbol{\alpha}_{CB} = \{8, 2\}$. We expect a mean 20% of EMB, as suggested by Bouy et al. (2015b). For fractions in the proper motions GMM, hyper-parameter are $\boldsymbol{\alpha}_{Cs} = \{1, 1, 5, 5\}$ and $\boldsymbol{\alpha}_{Bs} = \{1.2, 8.8\}$. These values induce fraction distributions whose means are similar to the fractions recovered after fitting a GMM to the Bouy et al. (2015b) candidate members. For the fraction in the GMM of

the CI distribution, the hyper-parameter were set all to 1, ($\boldsymbol{\alpha}_{CI} = \{1, 1, 1, 1, 1\}$), which results in equal means and large variances for all of them.

In the previous cases, with exception of the cluster-field mixture, the hyper parameters were chosen to have larger variances. Fig. 3.2 shows the distribution associated to these hyper-parameters.

The narrow variance in the cluster-field mixture expresses our prior belief about the number (fraction) of candidate members within our large data set. I believe this figure is small.

Figure 3.2: Prior distribution of fraction parameters. From top left to bottom right, the distributions of field fraction (π), equal mass binaries fraction ($1-\pi_{CB}$), and the cluster (π_{Cs}) and equal-mass binaries (π_{Bs}) fractions in their proper motion GMM, respectively.

For the priors of mean parameters in the proper motions GMM, both of single and EMB, we choose the bivariate normal distribution. We set the hyper-parameters of this bivariate normal to the MLE found after fitting a bivariate normal to the candidate members of Bouy et al. (2015b). These values are $\boldsymbol{\mu}_{\mu_{pm}} = \{16.30, -39.62\}$ and $\Sigma_{\mu_{pm}} = \{\{36.84, 1.18\}, \{1.18, 40.71\}\}$.

As prior for the covariance matrices of both single and EMB proper motions GMM we use the Half–$t(\nu, \mathbf{A})$ distribution. With ν an scalar and $\mathbf{A}$ a vector of dimension equal to that of the space. As shown by Huang & Wand (2013) this family distribution leads to more accurate estimation of covariance matrices than the traditional Inverse-Wishart distribution. Particularly, the marginal correlation correlation parameters, ρ have the following distribution,

$$p(\rho) \propto (1 - \rho^2)^{\frac{\nu}{2}-1} \tag{3.37}$$

while the standard deviation term σ_k associated to entry k is distributed according to Half–$t(\nu, A_k)$. We set the hyper-parameter to $\nu = 3$ and $\boldsymbol{A}_{pm} = \{10^5, 10^5\}$. According to Huang & Wand (2013), an arbitrarily large values of $\boldsymbol{A}$ lead to arbitrarily weakly informative priors on the corresponding standard deviation terms.

Photometric priors can be grouped in three categories, those concerning: (i) the *true* CI, (ii) the splines coefficients, and (iii) the cluster sequence intrinsic

dispersion.

For the means in the univariate GMM modelling CI we use a uniform in the range ($0.8 \leq CI \leq 8$). For the standard deviations I choose the Half–Cauchy$(0, \eta)$ distributions as suggested by Gelman (2006). I choose the arbitrarily large value of $\eta = 100$..

For the coefficients in the spline series we set the priors as univariate normal distributions. To find the values of the hyper-parameters, we proceed as follows. First, we remove EMB from Bouy et al. (2015b) candidate members. I performed an iterative fit of the cluster sequence, such that in each iteration objects above the 0.75 magnitude were removed. In the region of $CI > 7$ no candiadte members have been found. Thus, to provide a prior we complement our list of candidate members with the brown-dwarfs from the Faherty et al. (2012) sample. Only those with the same CMD as our data set. Finally, we fit the splines, and use the coefficients of this fit as means, μ_β of the univariate normal distributions. The standard deviation terms were set to $\sigma_\beta = \{1, 1, 1, 1, 1, 0.5, 0.1\}$. These values provide a reasonable compromise between cluster sequences compatible with the previously known candidates and those far away or with exotic shapes. We show a sample of this priors in Fig. 3.3. This Fig. also shows the brown-dwarfs from Faherty et al. (2012) and the sequence (dashed line) we use to provide the means of the univariate normal distributions.

To set the prior for the parameters of the cluster intrinsic dispersion, I choose again the Half–t$(\nu, \boldsymbol{A})$ distribution. However, this time I use $\boldsymbol{A}_{ph} = \{10, 10, 10, 10, 10\}$. These values are large when compared to the standard deviation terms o the observation uncertainty, thus provide a weakly informative prior on the marginal standard deviation terms of the Σ_{clus} covariance matrix.

Table **??** shows a summary all the hyper-parameter and their values.

Figure 3.3: CMD K_s vs. $i - K_s$ showing a sample of the prior for the coefficients in the splines series. Also shown the brown-dwarfs we add from Faherty et al. (2012) sample, and the cluster sequence (dashed line) found after fitting the splines to the brown-dwarfs and candidate members below the equal-mass binaries sequence.

3.5 Sampling the posterior distribution

Theoretically, there are at least three possible approaches to obtain the posterior distributions of the parameters in our model. One of these theoretical options is the analytical solution. This solution is obviously not feasible, due to our large data set. The second theoretical option consists of a grid in parameter space. The likelihood and the prior must be evaluated at each point in this grid and then multiplied. This approach is reasonable when the parametric space is of moderate dimension. It requires the evaluation of the posterior distribution q^p times, with q the number of grid points in one dimension, and p the dimension of the parametric space. The number of parameters in our model is 85, which immediately rules out this possibility. The third and so far only feasible approach is the use of Markov Chain Monte Carlo (MCMC). Although these methods provide a solution in a reasonable time, nevertheless, the bottle neck of computing time is due to the evaluation of the likelihood, which rows linearly with the size of the data set.

This Section is structured as follows. First, I introduce an heuristic technique to search for the maximum a posteriori of our target distribution: the posterior distribution of the parameters given the data. Then I will describe the MCMC techniques available in the literature. Particularly the one we chose and the reasons for which it was chosen. I will end this section with the details about the assessment of the MCMC convergence.

3.5.1 PSO

The likelihood of the data is the product of the individual likelihoods of each datum (Eq. **??**). Therefore, the number of operations needed to evaluate the likelihood grows proportionally to the size of the data set. As I will explain in Section **??**, the burn-in phase of MCMC techniques allows them to reach the target distribution. However, once the MCMC reaches the target distribution, the burn-in computations are discarded. Since the evaluation of the likelihood, and therefore of the posterior, is computationally expensive due to our large data set, I decided to reduce as much as possible the burn-in phase. To do so, I provide MCMC with a set of initial solutions which are close to the Maximum-A-Posteriori (MAP) of the target (posterior) distribution. These MAP solution must not be crowded on

the MAP solution, otherwise MCMC will spend a lot of time in expanding this initial set of solutions. Our objective is a representative sample of the full posterior distribution, not just an estimate of it.

In Section **??** I will show that the MCMC flavour more suitable to our objective belongs to the family of *ensemble* MCMC. This flavour works with particles in the parametric space. To make the transition between the initial MAP solutions and the particle MCMC as efficient as possible I chose a MAP finder which works also with particles. The Particle Swarm Optimiser (PSO, Kennedy & Eberhart 1995) provides a heuristic cheap and fast approach to the MAP solution. The PSO works with an ensemble of particles which move in through the parametric space. These particles use the collective and individual past and present information to update their position. This information is specified by the score function, which in our case is the posterior distribution. The particles update their position iteratively by means of a velocity. It has a random but restricted magnitude, however, its direction is determined by the object position and the individual and collective positions of maximum score. Kennedy & Eberhart (1995) has a detailed description of the original algorithm, while a more efficient version is given by (Clerc & Kennedy 2002).

Although the PSO is a simple and rather efficient solution to the MAP approximation, it is far from perfect. Due to its heuristic origin, there is no theory behind its formulation, furthermore, it does not always guarantee the finding of the global maxima. A convergence guaranteed version can be found in Patel et al. (2013). This issue does not affect our results because MCMC does guarantee the target distribution. However, it has an impact in the computing time. Also, PSO stops once the particles scores are within a user defined tolerance. If the tolerance is to large, the PSO may not converge. If the tolerance is small it may converge but deliver solutions highly concentrated around the solution. This poses a problem to the following MCMC stage. In order to explore the full posterior distribution MCMC needs more iterations to expand the initially concentrated positions. To overcome this problem I decided to use the charged PSO (Blackwell & Bentley 2002)

3.5.1.1 The charged PSO

Originally designed to optimise a time varying score function, the charged PSO maintain its exploratory capabilities due to an electrostatic force that repels particles when they got closer than a certain distance (Blackwell & Bentley 2002). Thanks to this electrostatic force the charged PSO avoids the over-crowding of particles around local best values.

The algorithm of Blackwell & Bentley (2002) computes distances in the entire parametric space. I find this approach unsuitable to our problem, thus I modified it. This modified version is described in more details on Section 3.6.1.

3.5.2 MCMC

3.5.2.1 Generalities

Markov chain Monte Carlo is the generic name for a series of algorithms whose objective is the sampling of probability distributions. As their name indicates, MCMC generate a chain of Monte Carlo realisations that fulfil the Markov property. Monte Carlo realisations can be understood, broadly speaking as continuous random realisations. Since it is an iterative algorithm, the chain refers to the joint of al random Monte Carlo steps. The Markov property refers to the probabilistic independence of the steps in the chain. A Markov chain is that in which the probability of a future step depends only on the present step, and not in the past steps.

Andrieu et al. (2003) provides a brief and interesting summary of the history of the MCMC methods. In the following I use their work to describe the fundaments of MCMC. For more details, see the aforementioned authors and the book of Brooks et al. (2011).

A stochastic process is defined as a sequence $\{\theta_1, ..., \theta_n\}$ of random elements from a set, where each element $\theta_i \subset \mathbb{R}^k$ with k the dimension of the *state space*.

A stochastic process, $\boldsymbol{\theta} = \{\theta_0, \theta_1, ..., \theta_n, \theta_{n+1}\}$ is called a Markov chain if

$$p(\theta_{n+1}|\theta_0, \theta_1, ...\theta_n) = p(\theta_{n+1}|\theta_n).$$

A Markov chain has two important distributions, the initial distribution and

the transition distribution. The initial distribution is the marginal distribution of θ_0, it is $p(\theta_0)$. The transition distribution is the conditional probability $p(\theta_{n+1}|\theta_n)$. This last one is called stationary or homogeneous if it does not depend on n.

If this transition is irreducible and aperiodic, then there is an invariant or equilibrium distribution to which the chain converge in spite of the initial distribution. Aperiodic means that the chain does not make loops. It is irreducible if the probability of exploring all other states is not zero.

If we want to have $p(\theta)$ as the invariant distribution, then it suffices that the transition distribution $p_t(\cdot|\cdot)$ satisfies the detailed balance condition,

$$p(\theta_n) \cdot p_t(\theta_{n-1}|\theta_n) = p(\theta_{n-1}) \cdot p_t(\theta_n|\theta_{n-1}) \tag{3.38}$$

Thus MCMC are Markov chains that satisfy the detailed balance and had their invariant distribution as the target distribution. The variety of MCMC algorithms arises from the efficiencies in which they arrive to the target distribution.

In the following I will review three of the MCMC categories: Metropolis-Hasting (MH), Hamiltonian Monte Carlo (HMC), and affine invariant samplers. The MH category comprises the classic MH algorithm but also contain particular cases like the Gibbs sampler (Geman & Geman 1984). I describe MH only for completeness and explanatory reasons. Later, I will focus on the particular cases of Hamiltonian Monte Carlo, for (HMC), affine invariant for ensemble samplers. Finally, I will briefly describe Nested Sampling, an algorithm that uses MCMC to numerically compute the Bayesian evidence while simultaneously samples the posterior distribution.

3.5.2.2 Metropolis-Hastings

By far, the most popular MCMC algorithm is Metropolis-Hastings (Metropolis et al. 1953; Hastings 1970). Given the current, θ, and proposed, $\hat{\theta}$ positions of the Markov chain, which live in the state space, the chain moves from θ to $\hat{\theta}$ with acceptance probability

$$\mathcal{A}(\hat{\theta}|\theta) = min\left\{1, \frac{p(\hat{\theta}) \cdot q(\theta|\hat{\theta})}{p(\theta) \cdot g(\hat{\theta}|\theta)}\right\}. \tag{3.39}$$

Where q is the transition probability. Since the algorithm allows rejection, it is aperiodic, and to ensure irreducibility, the support of q must include that of p (Andrieu et al. 2003). The popularity of MH lies in its simplicity. Nevertheless it requires a careful tuning of the transition probability.

3.5.2.3 Hamiltonian Monte Calro

The Hamiltonian Monte Carlo algorithms (**??**), as they name suggests[12], use Hamiltonian dynamics to express the target distribution as the potential distribution of a hamiltonian system. In such systems total energy is the sum of the potential and kinetic energies. The potential distribution depends only on position, whereas the kinetic one on momentum. To this end, HMC introduces a momentum to fictitious particles to use their positions as a sample of the target distribution. To update the particles positions, HMC uses the Hamilton equations. They contain information about the gradient of the potential. Once HMC has tuned the momentum distribution, the proposed positions are more likely in terms of the target distribution. Therefore, using the information about the gradient of the target distribution, HMC is able to improve the acceptance ratio of the proposed steps. A detailed description of HMC can be found in Chapter 5 of Brooks et al. (2011). The package *Stan* (**?**) provides an efficient implementation of HMC.

3.5.2.4 Affine invariant

Affine invariant MCMC samplers use many particles, (ensemble), to sample the target distribution with a performance that is independent of its shape in the parametric space. Affine invariant MCMC do not need to tune the transition distribution, for this reason, these samplers are faster than standard MCMC (Goodman & Weare 2010). In the following I use the derivation of Goodman & Weare (2010).

An ensemble $\boldsymbol{\theta}$ is a set of L particles $\theta_l \subset \mathbb{R}^k$ living in state space $\mathbb{R}^{kL}$. These particles are independently drawn from the target distribution π. Therefore,

$$\Pi(\boldsymbol{\theta}) = \pi(\theta_1) \cdot \pi(\theta_2)...\pi(\theta_L).$$

Thus, an ensemble MCMC is a Markov chain in the state space of ensembles,

[12]Originally called Hybrid Monte Carlo by (**?**)

more properly, in the state space of the sequence $\boldsymbol{\theta}(1), \boldsymbol{\theta}_(2), ...\boldsymbol{\theta}(t)$. An ensemble MCMC preserves the equilibrium distribution without the individual particles sequence, $\theta_1(1), \theta_1(2), ...\theta_1(t)$, being Markov or even independent.

To update the particles positions, the detailed balance must be fulfilled. Goodman & Weare (2010) use partial resampling to ensure this. The transition preserves the target distribution if the single particle move preserves the conditional distribution of the particle given the complementary ensemble (the rest of the particles). Using the affine invariant *stretch move*, they are able to define a Markov chain (in the state space of ensembles) that satisfies the detailed balance. The stretch move $\theta_k(t) \rightarrow \hat{\theta}$, defined as,

$$\hat{\theta} = \theta_j(t) + z \cdot (\theta_k(t) - \theta_j(t)),$$

where $\theta_j(t)$ is the current position of a particle in the complementary ensemble, and z is the stretching factor. It produces a symmetric transition, $p(\theta_k(t) \rightarrow \hat{\theta}) = p(\theta_k(t) \leftarrow \hat{\theta})$, if its density $g(z)$ satisfies the symmetry condition

$$g(\frac{1}{z}) = z \cdot g(z).$$

Finally, Goodman & Weare (2010) define their affine invariant MCMC using $g(z) \propto 1/z, z \in [1/a, a]$ and zero otherwise, together with acceptance probability,

$$\mathcal{A}(\hat{\theta}|\theta) = min \left\{ 1, z^{n-1} \cdot \frac{p(\hat{\theta})}{p(\theta)} \right\}. \tag{3.40}$$

The parameter $a > 1$ improve the performance the performance of the sampler (Goodman & Weare 2010).

One of the greatest advantages of ensemble samplers is its possibility of parallelisation. Since they work with particles, this particles can be distributed among cores in a computer cluster, therefore reducing considerably the computing time. Foreman-Mackey et al. (2013) implemented the affine invariant stretch move of Goodman & Weare (2010) in the Python package *emcee*.

3.5.2.5 Nested sampling

Nested sampling (**??**) is an algorithm designed to numerically integrate the evidence (Eq. 3.12). As a by product it delivers also a sample of the posterior

distribution. It samples the prior and use these particles to compute integral. The integral is the sum of the likelihood, L_i of each particle times a weight, w_i. This weight is a proxy of the volume of the prior covered by the updated position of the particle, $w_i = \exp(-(i-1)/N) - \exp(-i/N)$. Particles update their position only when their likelihood is the minimum from the group of particles. In this way,

$$z \leftarrow \sum_i w_i \cdot L_i \tag{3.41}$$

An improved version of the original algorithm was implemented in *MultiNest* (**?**). This version allows the sampling and computing of evidence in multimodal posteriors.

3.5.2.6 Implementation and convergence

To sample the posterior distribution in our problem, we chose *emcee* due to the following properties: i) the affine invariance allows a faster convergence over common and skewed distributions (see Goodman & Weare 2010; Foreman-Mackey et al. 2013, for detail), ii) the parallel computation distributes particles over nodes of a computer cluster and thus reduces considerably the computing time, and iii) it requires the hand-tuning of only two constants: the number of particles, and the parameter a of the $g(z)$ distribution. I choose a particle to parameters ratio o two, this is 170 particles. This is the minimum recommended by Foreman-Mackey et al. (2013) that allows a reasonable computing time. After trial and error, I fixed the value of parameter $a = 1.3$. As mentioned by Goodman & Weare (2010) this parameter can be tuned to improve performance of the sampler. The chosen value keeps the acceptance fraction in the range $0.2 - 0.5$ as suggested by Foreman-Mackey et al. (2013).

I use a modified version of *CosmoHammer* (Akeret et al. 2013), a front-end of *emcee*. This enables us to control the input and output of data and parameters, as well as the hybrid parallel computing. For this last one, instead of using OpenMP as (Akeret et al. 2013) did, we use the *multiprocessing* package of python. It distributes the computing of the likelihood among cores in each cluster node. I run this code on a 80 CPUs (cores) computer cluster with 3.5 GHz processors.

As mentioned earlier, the PSO does not guarantee the finding of the global

maximum of the score function. Therefore, I decide to implement an iterative approach that minimises the risk of PSO to get stuck in a local maxima. To do so I iteratively run PSO and 50 iterations of *emcee* (with the same number of particles as the PSO) until the relative difference between means of consecutive iterations was lower than 10^{-7}. The iterations of *emcee* spread the PSO solution without moving away from the maximum.

Neither scheme, PSO alone or PSO-*emcee*, guarantees to find the global maximum and their solution could be biased. However, we use them to obtain a fast estimate of the global maximum, or at least, of points in its vicinity. If the initial solution provided by this scheme is indeed biased, the final *emcee* run, during the burning phase, erases any dependance on these initial solutions. After convergence of the PSO-*emcee* scheme, we run *emcee* until convergence.

Convergence to the target distribution occurs when each parameter enters into the stationary equilibrium, or normal state. The Central Limit Theorem ensures that this state exists. See Roberts & Rosenthal (2004) for guaranteeing conditions and Goodman & Weare (2010) for *irreducibility* of the *emcee* stretch move. The stationary or normal state is reached when, in at least 95% of the iterations, the sample mean is bounded by two standard deviations of the sample, and the variance by the two standard deviation of the variance [13]; see Fig. 3.4.

Figure 3.4: Normalised mean (left panel) and variance (right panel) of each parameter in our model, as functions of iterations. The normalisation values are the mean and variance of the ensemble of particles positions at the last iteration. Red lines show one and two sigma levels of these normalisation values.

Once all parameters have entered the equilibrium state, we stop *emcee* sampling using the criterium of Gong & Flegal (2016) [14]. We chose this criterium because it was developed for high-dimensional problems and tested on Hierarchical Bayesian Models. In this criterium, the MCMC chain stops once its "effective sample size" (ESS, the size that an independent and identically distributed sample must have to provide the same inference) is larger than a minimum sample size computed

[13] $sd(\sigma^2) = \sigma^2\sqrt{\kappa/n + 2/(n-1)}$ with κ the kurtosis and n the sample size.

[14] Implemented in the R package *mcmcse* (Flegal et al. 2016)

using the required accuracy, υ, for each parameter confidence interval $(1-\delta)100\%$. Our *emcee* run stops once the ESS of the ensemble of walkers is greater than the minimum sample size needed for the required accuracy $\epsilon = 0.05$ on the 68% confidence interval ($\delta = 0.32$) of each parameter.

3.6 Codes

This Section sets out the details about the code I develop to perform the computation described throughout this chapter. First, I give a brief chronological description of the model development. Later, I will describe the details on the implementation of:, the charged PSO, the modified *emcee*, and the GMM used to describe the field population. Finally, I will end the this Section detailing the hybrid high performance computing code developed to minimise the computing time of the BHM.

The first version of the Bayesian Hierarchical Model was implemented by Ángel Berihuete in the package Stan (). It comprised a Bayesian model of the ML model of Sarro et al. (2014). The proper motions were modelled using a single mixture of gaussians. The photometry was modelled with a Chebyshev polynomial parametrised with the length along the sequence. This length was found using a principal curve analysis. Later, I included the EMB and uncertainties both in proper motion and in photometry, and the width of the sequence modelled as the multivariate gaussian. We realised that the principal curve analysis is not compatible with the deconvolution methodology. The principal curve analysis is driven by the elements with higher variance. We decided to parametrise the polynomials with the true colour, a nuisance parameter that was marginalised with the aid of a prior. For this prior we introduced the colour distribution. Previous to the introduction of the marginalisation of the nuisance parameters, the model worked fine on samples of a few thousands of stars. One the marginalisation was introduced the model increased the amount of computing time rendering it useless for higher data sizes. At this point we decided to port the existing code into python in order to use the parallel *emcee*. *emcee* proved to be of greater use. Due to its parallelisation capabilities we were able to increase the data size from 2000 to 10,000 objects. Since the computing of the likelihood was the highest computa-

tional challenge, I developed my own routines to perform it on parallel (see Section **??** below). However, *CosmoHammer* (Akeret et al. 2013) turn out to be more efficient in distributing the loads. I ported the code into *CosmoHammer* and modified this last one to better suit our needs. Despite the Hybrid-HPC, the computing of the likelihood of 10^5 seemed unreachable. At this point I performed two tasks, the first was to strip the code of all auxiliary libraries calls, and the vectorisation of the operations. Since the parameters of the field were held fixed, the field likelihood was computed externally for each object. The code was then fed with the data, the field likelihood, and the auxiliary computations reduced at minimum. These reduction included for example, the Cholesky decompositions and matrix inversions of the uncertainty. Thus, instead of doing these computation inside the code, the code was fed with the appropriate values.

Introducing PSO and later the charged PSO reduced the computing time. At this point we were able to compute the first 10^5 run, which however did not converged in reasonable time. Once the approximation of the marginalisation integral (Eq. **??**) was introduced the computing time reduced far more. Finally, the tuning of the parameter allowed us to increase the acceptance fraction, and reach convergence within 4 weeks of full computing time in a 80 cores computer cluster. It is indeed a very long time, however it is reasonable compared with our original estimates of approximately 2 years[15].

3.6.1 The modified charged PSO

As explained before, the charged PSO of Blackwell & Bentley (2002) was inappropriate to our objective. The metric of the parametric space of our problem is not isotropic, parameters have different length scales. For example, while fraction are constrained in the $[0, 1]$ interval proper motions parameters can go far larger than that. Therefore, the use of an isotropic metric results in a solution which is crowded in some parameters while is over dispersed in others. To solve this issue, I modified the charged PSO by measuring distance between particles and applying the electrostatic force independently in each parameter. In such a way, the electrostatic force plays a role only when the relative distance between particles is

[15]Today, the DANCe team is working on a GPU version of the code which is expected to compute the same amount of calculations in a couple of days.

smaller than 10^{-10}. I found this value heuristically.

In the original version of Blackwell & Bentley (2002), each particle is subject to the acceleration,

$$\mathbf{a} = \sum_{i \neq j} \frac{q_i \cdot q_j}{r_{ij}^3} \cdot \mathbf{r}_{ij}, \qquad p_{core} < r_{ij} < p \tag{3.42}$$

where q_i and q_j are the charges of particles i and j, r_{ij} is the distance between them. The distances p_{core} and p indicate the minimum and maximum distances at which the electrostatic force came into action. Outside this range, the electrostatic force is zero. In this equation, $\mathbf{r}_{ij} = \mathbf{x}_i - \mathbf{x}_j$, where $\mathbf{x}_i, \mathbf{x}_j$ are the positions of particles i and j. Also, $\mathbf{r}_{ij}, \mathbf{x}_i, \mathbf{x}_j \subset \mathbb{R}^d$, with d the dimension of the space.

In the modified version, the distance is measured independently in each dimension of the parametric space. Thus, $\mathbf{r}_{ij} = \{x_{1,i} - x_{1,j}, x_{2,i} - x_{2,j}, ..., x_{d,i} - x_{d,j}\}$. Also the acceleration has the form,

$$\mathbf{a} = \sum_{i \neq j} \frac{q_i \cdot q_j}{r_{ij}^2} \cdot \mathbf{r}_{ij}, \qquad 10^{-50} < \frac{r_{ij}}{r_{eq}} < \epsilon \tag{3.43}$$

and it is now applied over each dimension of the parametric space. The distance r_{eq} is that at which the velocity caused by the acceleration equals the mean velocity caused by the common PSO. ϵ is a free parameter set, heuristically, to 10^{-10}.

3.6.2 Improvements of emcee

The modification I introduced in *emcee*, although very simple, improved the acceptance fraction and mixing of the particles. To allow the parallelisation, Foreman-Mackey et al. (2013) divide the ensemble of particles in two ensembles. In the original version, the particles in one ensemble use one and the same particle in complementary ensemble to compute their positions according to Eq. 3.40. In the modified versions, particles from one ensemble update their positions using a particle from the complementary ensemble, however, this particle is chosen randomly at each iteration.

Figure **??** compares the mixing and acceptance fractions of the original and modified versions of *emcee*. In a private communication with David Foreman, he

mentioned that a similar modification was already introduced in a beta version of *emcee*.

3.6.3 GMM for the field population

3.6.4 Parallel implementations

However, since MPI (Message Passing Interface) was already used by *emcee* to distribute the the particles across cores. I realised a hybrid approach was necessary. In this hybrid HPC approach, *emcee* particles were distributed in nodes while objects were distributed across cores in a node. Description of the implementation. MPI, python stan, etc. explain in detail the difficulties faced at implementing the different codes in the different servers.

Chapter 4

Results

4.1 Performance of the classifier

Compare and explain the differences with Stauffer members.

4.2 Velocity distribution

4.3 Spatial distribution

4.4 Luminosity distribution

4.5 Mass distribution

4.5.1 The mass-luminosity relation

4.6 The mass distribution on time

4.7 The phase space

4.8 Updating the previous knowledge

Chapter 5

Conclusions and Future Work

Appendix A

Appendix Chapter Title

In the following we include ....

Bibliography

????

Adams, J. D., Stauffer, J. R., Monet, D. G., Skrutskie, M. F., & Beichman, C. A. 2001, AJ, 121, 2053

Akeret, J., Seehars, S., Amara, A., Refregier, A., & Csillaghy, A. 2013, Astronomy and Computing, 2, 27

Allard, F., Homeier, D., Freytag, B., & Sharp, C. M. 2012, in EAS Publications Series, Vol. 57, EAS Publications Series, ed. C. Reylé, C. Charbonnel, & M. Schultheis, 3–43

Alsing, J., Heavens, A., & Jaffe, A. H. 2017, MNRAS, 466, 3272

Anderes, E., Wandelt, B. D., & Lavaux, G. 2015, ApJ, 808, 152

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. 2003, Machine Learning, 50, 5

Azzalini, A. 1996, Biometrika, 83, 715

Azzalini, A. & Capitanio, A. 1999, Journal of the Royal Statistical Society: Series B (Methodological)

Balaguer-Núñez, L., Galadí-Enríquez, D., & Jordi, C. 2007, A&A, 470, 585

Barnes, III, T. G., Moffett, T. J., Jefferys, W. H., & Forestell, A. D. 2004, in Astronomical Society of the Pacific Conference Series, Vol. 310, IAU Colloq. 193: Variable Stars in the Local Group, ed. D. W. Kurtz & K. R. Pollard, 95

Barrado, D., Bouy, H., Bouvier, J., et al. 2016, A&A, 596, A113

Bishop, C. 2006, Pattern Recognition and Machine Learning, Information Science and Statistics (Springer)

Blackwell, T. & Bentley, P. 2002, in Proceedeings of the 4th Annual Conference on Genetic and Evolutionary Computations, ed. W. Langdon, E. Cantu-Paz, K. Mathias, R. Roy, & D. Davis (San Francisco, CA, USA: Morgan kaufmann Publishers Inc.), 19–26

Bouvier, J., Kendall, T., Meeus, G., et al. 2008, A&A, 481, 661

Bouy, H., Bertin, E., Barrado, D., et al. 2015a, A&A, 575, A120

Bouy, H., Bertin, E., Moraux, E., et al. 2013, A&A, 554, A101

Bouy, H., Bertin, E., Sarro, L. M., et al. 2015b, A&A, 577, A148

Bovy, J., Hogg, D. W., & Roweis, S. T. 2011, The Annals of Applied Statistics, 5, 1657

Brooks, S., Gelman, A., Jones, G., & Meng, X. 2011, Handbook of Markov Chain Monte Carlo, Chapman & Hall/CRC Handbooks of Modern Statistical Methods (CRC Press)

Buchner, J., Georgakakis, A., Nandra, K., et al. 2014, A&A, 564, A125

Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, ApJ, 345, 245

Chabrier, G. 2003, PASP, 115, 763

Chabrier, G. 2005, in Astrophysics and Space Science Library, Vol. 327, The Initial Mass Function 50 Years Later, ed. E. Corbelli, F. Palla, & H. Zinnecker, 41

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. 2015, Journal of Educational and Behavioral Statistics, 40, 136

Clerc, M. & Kennedy, J. 2002, Trans. Evol. Comp, 6, 58

Converse, J. M. & Stahler, S. W. 2008, ApJ, 678, 431

D'Antona, F. 1998, in Astronomical Society of the Pacific Conference Series, Vol. 142, The Stellar Initial Mass Function (38th Herstmonceux Conference), ed. G. Gilmore & D. Howell, 157

De Boor, C. 1978, A practical guide to splines / Carl de Boor (Springer-Verlag New York), xxiv, 392 p. :

Demory, B.-O. 2014, ApJ, 789, L20

Diard, J. & Bessiere, P. 2008, Probabilistic reasoning and decision making in sensory-motor systems, 153

Dirac, P. A. M. 1942, Proceedings of the Royal Society of London Series A, 180, 1

Duquennoy, A. & Mayor, M. 1991, A&A, 248, 485

Faherty, J. K., Burgasser, A. J., Walter, F. M., et al. 2012, ApJ, 752, 56

Fan, Y. & Sisson, S. 2011, Handbook of Markov Chain Monte Carlo, 67

Feeney, S. M., Johnson, M. C., McEwen, J. D., Mortlock, D. J., & Peiris, H. V. 2013, Phys. Rev. D, 88, 043012

Fei-Fei, L. & Perona, P. 2005, in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 2, Ieee, 524–531

Flegal, J. M., Hughes, J., & Vats, D. 2016, mcmcse: Monte Carlo Standard Errors for MCMC, Riverside, CA and Minneapolis, MN, r package version 1.2-1

for Guides in Metrology, J. C. 2008, JCGM 100: Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement, Tech. rep., JCGM

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, Publications of the Astronomical Society of the Pacific, 125, 306

Fraley, C. & Raftery, A. E. 2002, Journal of the American Statistical Association, 97, 611

Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. 2012, mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation

Franklin, J. 2002, The Science of Conjecture: Evidence and Probability Before Pascal, A Johns Hopkins paperback (Johns Hopkins University Press)

Gagné, J., Lafrenière, D., Doyon, R., Malo, L., & Artigau, É. 2014, ApJ, 783, 121

Galli, P. A. B., Moraux, E., Bouy, H., et al. 2017, A&A, 598, A48

Gelman, A. 2006, Bayesian Analysis, 1, 515

Gelman, A. 2012, CHANCE, 25, 52

Gelman, A., Carlin, J., Stern, H., et al. 2013, Bayesian Data Analysis, third edition edn. (CRC)

Gelman, A. & Hill, J. 2007, Data Analysis Using Regression and Multilevel/Hierarchical Models, Analytical Methods for Social Research (Cambridge University Press)

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. 2008, Ann. Appl. Stat., 2, 1360

Geman, S. & Geman, D. 1984, IEEE Trans. Pattern Anal. Mach. Intell., 6, 721

Gong, L. & Flegal, J. M. 2016, Journal of Computational and Graphical Statistics, 25, 684

González-Farias, G., Dominguez-Molina, A., & Gupta, A. K. 2004, Journal of Statistical Planning and Inference, 126, 521

Good, I. J. 1980, Trabajos de Estadistica Y de Investigacion Operativa, 31, 489

Goodman, J. & Weare, J. 2010, Communications in Applied Mathematics and Computational Science, 5, 65

Gupta, A. K., González-Farias, G., & Dominguez-Molina, J. 2004, Journal of Multivariate Analysis, 89, 181

Guthrie, B. N. G. 1987, QJRAS, 28, 289

Hand, D. J. & Yu, K. 2001, International Statistical Review / Revue Internationale de Statistique, 69, 385

Hastings, W. K. 1970, Biometrika, 57, 97

Hogg, D. W., Bovy, J., & Lang, D. 2010a, ArXiv e-prints

Hogg, D. W., Myers, A. D., & Bovy, J. 2010b, ApJ, 725, 2166

Hong, D., Balzano, L., & Fessler, J. A. 2016, Towards a theoretical analysis of PCA for heteroscedastic data, arxiv 1610.03595

Huang, A. & Wand, M. P. 2013, Bayesian Analysis, 8, 439

Jaynes, E. 2003, Probability Theory: The Logic of Science (Cambridge University Press)

Jefferys, T. R., Jefferys, W. H., Barnes, III, T. G., & Dambis, A. 2007, in Astronomical Society of the Pacific Conference Series, Vol. 371, Statistical Challenges in Modern Astronomy IV, ed. G. J. Babu & E. D. Feigelson, 433

Kass, R. E. & Raftery, A. E. 1995, Journal of the American Statistical Association, 90, 773

Kennedy, J. & Eberhart, R. 1995, in Neural Networks, 1995. Proceedings., IEEE International Conference on, Vol. 4, 1942–1948 vol.4

Koller, D. & Friedman, N. 2009, Probabilistic Graphical Models: Principles and Techniques, Adaptive computation and machine learning (MIT Press)

Krone-Martins, A. & Moitinho, A. 2014, A&A, 561, A57

Lindstrom, M. 1999, Journal of Computational and Graphical Statistics, 8, 333

Liu, Z., Chen, Y., Tian, S., & Xu, Z. 2015, Journal of Information and Computational Science, 12, 775

Lodieu, N., Deacon, N. R., & Hambly, N. C. 2012, MNRAS, 422, 1495

Malo, L., Doyon, R., Lafrenière, D., et al. 2013, ApJ, 762, 88

March, M. C., Karpenka, N. V., Feroz, F., & Hobson, M. P. 2014, MNRAS, 437, 3298

McMichael, D. W. 1996, in in Proc. Fourth International Symposium on Signal Processing and its Applications (ISSPA, 377–378

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, The Journal of Chemical Physics, 21, 1087

Moraux, E., Bouvier, J., & Stauffer, J. R. 2001, A&A, 367, 211

Morris, C. N. 1983, Journal of the American Statistical Association, 78, 47

Muench, A. A., Lada, E. A., Lada, C. J., & Alves, J. 2002, ApJ, 573, 366

Olivares, J. & et al. 2017a, submitted to å

Olivares, J. & et al. 2017b, in preparation

Ozerov, A., Lagrange, M., & Vincent, E. 2013, Computer Speech and Language, 27, 874 , special Issue on Speech Separation and Recognition in Multisource Environments

Patel, P. K., Sharma, V., & Gupta, K. 2013, International Journal of Computer Applications, 73

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Press, W. H. 1997, in Unsolved Problems in Astrophysics, ed. J. N. Bahcall & J. P. Ostriker, 49–60

Pugachev, V. 1965, Theory of Random Functions and its Application to Control Problems (Pergamon), 852

Riedel, A. R., Blunt, S. C., Lambrides, E. L., et al. 2017, AJ, 153, 95

Roberts, G. O. & Rosenthal, J. S. 2004, Probability Surveys, 1, 20

Sale, S. E. 2012, MNRAS, 427, 2119

Sampedro, L. & Alfaro, E. J. 2016, MNRAS, 457, 3949

Sarro, L. M., Bouy, H., Berihuete, A., et al. 2014, Astronomy & Astrophysics, 14

Schwarz, G. 1978, Ann. Statist., 6, 461

Shariff, H., Dhawan, S., Jiao, X., et al. 2016, MNRAS, 463, 4311

Shkedy, Z., Decin, L., Molenberghs, G., & Aerts, C. 2007, MNRAS, 377, 120

Smyth, G. K. 1998, in Encyclopedia of Biostatistics, ed. Armitage, P. & Colton, T. (John Wiley and Sons)

Spiriti, S., Eubank, R., Smith, P. W., & Young, D. 2013, Journal of Statistical Computation and Simulation, 83, 1020

Stauffer, J. R., Hartmann, L. W., Fazio, G. G., et al. 2007, ApJS, 172, 663

Stauffer, J. R., Schultz, G., & Kirkpatrick, J. D. 1998, ApJ, 499, L199

Stoneking, C. J. 2014

Takeda, Y., Hashimoto, O., & Honda, S. 2017, PASJ, 69, 1

Tapiador, D., Berihuete, A., Sarro, L. M., Julbe, F., & Huedo, E. 2017, Astronomy and Computing, 19, 1

Thies, I. & Kroupa, P. 2007, ApJ, 671, 767

Trumpler, R. J. 1921, Lick Observatory Bulletin, 10, 110

Wolfgang, A. & Lopez, E. 2015, ApJ, 806, 183